Google Colab Notebook- How to Use

Press and hold shift, then press enter to run a cell. Basically you need to run each cell one after the other to get the notebook to run. The entire notebook is linked to the Google sheet that I showed in the call- that sheet will need to stay in the same structure as it currently is for the notebook to work.

When you run the first couple of cells you'll be prompted to follow a link, copy a security code and paste it into the box you see in the notebook. This is just some Google security stuff.
The first section of analysis looks at wage data via multivariate regression- you can look at the results here but they won't be significant until you get a lot more data I'd say.

If someone is interested in understanding the regressions results a bit more- they can Google "interpreting statsmodels regression summaries" or something like that.
https://www.pluralsight.com/guides/interpreting-data-using-statistical-models-python
That article gives a good overview of what the different values in the tables mean.
StackOverflow is also a really good website for finding the answers to data questions.

Logistic Regression
The section of the notebook on logistic regression is probably going to be more useful to you. It looks at what factors affect the likelihood of someone graduating from the programme.
A lot of columns from your original sheet were collapsed/merged also. This is because the fewer categories there are for the data to be divided into, the more data there will be per category and so the higher the chance of statistically significant results being obtained.

The table below the cell titled "select the significant variables" outputs all the variables in the final logistic model that had a statistically significant effect on the likelihood of a person graduating.

There's a bunch of different cell outputs which have formatted the results nicely. The actual results are highlighted in magenta. They compare the probability of graduating if you are someone with 0% or no independent variable with the probability of graduating if you are someone with 100% or who has the independent variable.

This is perhaps a little confusing to get your head around. In the case of government assistance being the independent variable being measured, because of how the data

was collapsed you can only be classified as *receiving government assistance* or *not receiving government assistance.*
In the case of a continuous variable (something that has a number rather than a yes or no/category), then the regression is comparing someone who has 0 or 0% of it, versus someone who has the maximum/100% of it.

The final graph at the bottom is a graph comparing the probability of someone graduating from the programme if they have/have 100% of the independent variable or have no/0% of it.

Variable toggles:
Below the cell that says logistic regression there is a toggle for ethnicity. If you want ethnicity variables to be considered then leave it as it is- with the word yes typed in between two speech marks.
If you want to exclude ethnicity from the regression then replace yes with no.
Ethnicity may have an effect- but with the small sample size in the dataset we used, it didn't give any statistically significant results.

```
# select with or without ethnicity
ethnicity_included = 'yes'
```

In the cell below the section titled "look at model significance" there is a toggle that lets you pick which significant variable you want to look at in the graph.
Based on the results from when I ran the analysis, it can be filled with either 'data_completeness' or 'government_assistance'.
Any variable that appears in the table 2 cells above it can be entered in this toggle and the graph will change to reflect that, you just need to rerun all the cells between it and the graph again.

```
[ ] # Select variable you want to evaluate
    variable = 'data_completeness'
```

Results

Statistical significance- If a result has statistical significance then it means that the result was not just due to random chance. It is important that results have statistical significance if we are going to act on them. In the regressions, we say that any result with a p-value of less than 0.05 is statistically significant. You don't need to worry about p-values, but if you're curious- they are called P > |z| or P > |t| in the regression summary tables.

Model 1: Multivariate Linear Regression

This model looks at what factors affect the post-graduation wages that candidates received.
Because there is limited wage data currently available, no statistically significant results were obtained here.

```
[ ]  # Select variables where p < 0.1
     lm_coefficients[lm_coefficients['P>|t|'] <= 0.1]
```

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]

A p-value threshold of 0.1 was used here instead of 0.05 to see if results were even coming close to being statistically significant.
With more data, this regression may become more useful.

Model 2: Logistic Regression

This model looks at what factors affect the probability of someone graduating from the programme. The section below shows which variables had statistically significant effects on the probability of someone graduating from the programme.
The 'const' value can be ignored, it just refers to a value produced by the model.

Select the signficiant variables

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-0.837183	0.420342	-1.991671	4.640716e-02	-1.661039	-0.013328
government_assistance	-0.479341	0.209423	-2.288860	2.208749e-02	-0.889803	-0.068879
data_completeness	0.015902	0.002826	5.627711	1.826166e-08	0.010364	0.021440

The section below investigates the significant variables found from the table above in further detail.
It then outputs the different probabilities of graduating for that variable, depending on if you have it/100% of it or don't have it/have 0% of it.

```
[ ] # Select variable you want to evaluate
    variable = 'data_completeness'
```

Convert to probabilities and get variables for graph

Probability of graduating with variable

```
The probability of graduating when a candidate has a data_completeness (if categorical)
or has 100% data_completeness if it's a metric, is 75.36%
```

Probability of graduating without variable

```
The probability of graduating when a candidate doesn not have a data_completeness (if categorical)
or has 0% data_completeness if it's a metric, is 38.03%
```

Difference

```
When data_completeness is added in, after for controlling for
 ['gender', 'age', 'single_parent', 'government_assistance', 'college_educated', 'ethnicity_black'],
 inclusion of "data_completeness" means a person is 37.328% more likely to graduate
```

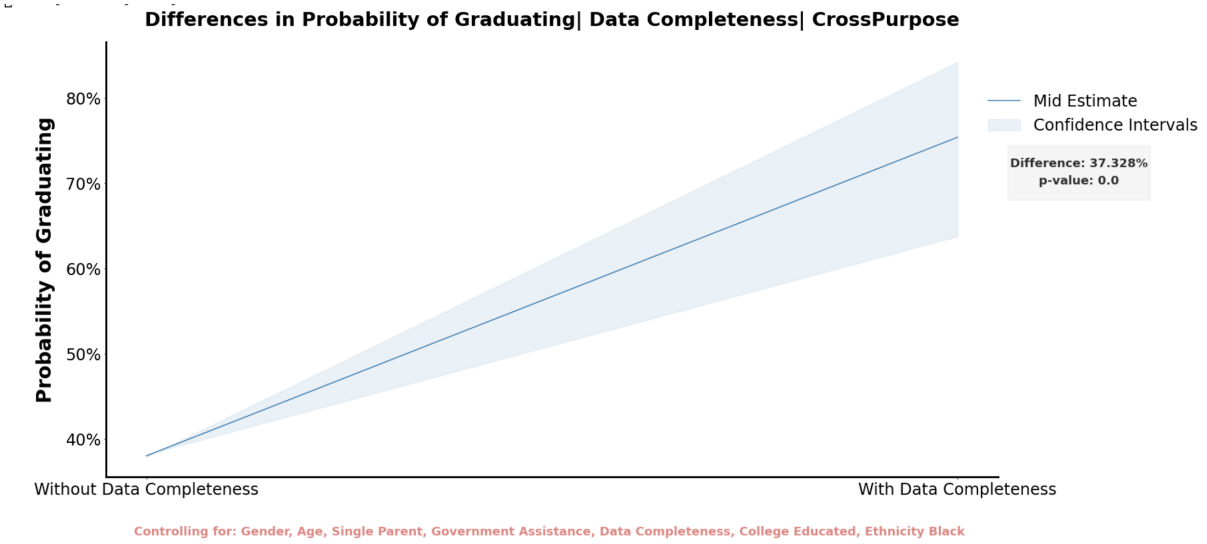
The final graph at the bottom just produces a visualization for the results above. The line represents the middle estimates of probabilities, and the shaded blue areas represent the confidence intervals of those results- an upper and lower bound for where the results could lie.
The section in red at the bottom of the graph tells you what significant variables are being controlled for in the model. It basically calculates who the 'average' person in your programme is - average age, most common gender, most common level of education etc- then compares the probability of this 'average' candidate graduating based on the independent variable. The average of a categorical variable like single parent or being black is just whichever the dataset has more people fall into. So the average of ethnicity black would be "not black" since more people are not black than are in this dataset.

So for this graph and this selected variable (using the magenta results from the cells above as well)-

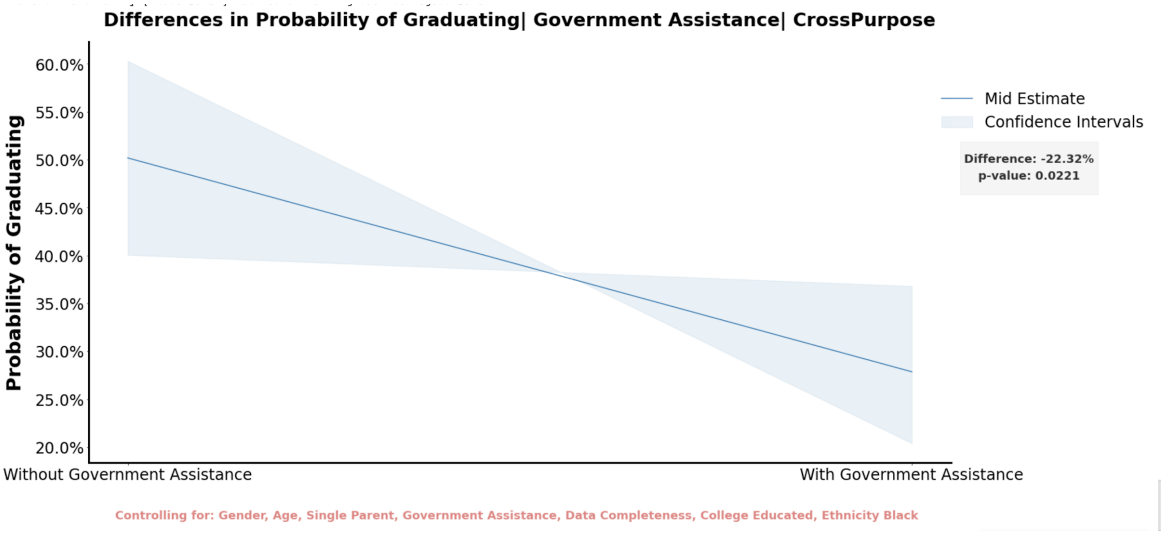
A person who had 100% data completeness (Had completed all the tests/assessments) was 75.36% likely to graduate and a person who had 0% data completeness was 38.03% likely to graduate. This was a difference of 37.33% as is seen in the graph.

The result was statistically significant ($p<0.05$), as is seen by the p-value in the graph as well.

This result controls for gender, age, single parent status, data completeness, being black, being on government assistance and education level.



The only other statistically significant result found was whether or not a candidate was receiving government assistance. Receiving government assistance meant the average candidate was 22.32% less likely to graduate from the programme. This was after controlling for all the variables in red at the bottom of the graph.



Top-line Summary

- The limited data meant that many variables/features of the dataset had to be combined- even still, few variables produced statistically significant effects (effects that we can be sure weren't just due to random chance).
- Without considerably more wage data, not much can be said about what factors affect it.
- The average candidate who had completed 100% of the assessment tests (Computer skills, Excel, social media etc) was 37.33% more likely to graduate than the average candidate who completed none of the tests. This result was statistically significant.
- The average candidate who was receiving some form of government assistance was 22.32% less likely to graduate than an average candidate who was receiving no form of government assistance. This result was statistically significant.